10_arxiv_(link analysis)

Marenco Turi Gualtiero, Ardigò Susanna

April 14, 2025

1 Introduction

This report presents a network analysis of the Arxiv HEP-PH (High Energy Physics Phenomenology) citation dataset. The dataset consists of papers published between January 1993 and April 2003 with directed edges representing citations: a directed edge from node i to node j means that paper i cites paper j. The goal of this study is to analyze and compare centrality measures to rank influential papers in the network.

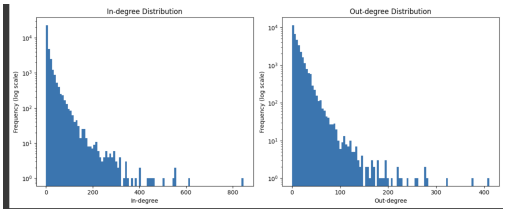
2 Dataset Description and Preprocessing

The dataset was provided in a text file format with over 34000 papers and more than 420000 citation links. The data was loaded using a custom Python function, and a directed graph was constructed using the `networkx` library. Summary statistics of citation counts (in-degree and out-degree) were computed for each node.

Figure 1: Summary Statistics of Degrees

	in_degree	out_degree
count	34546.000000	34546.000000
mean	12.203381	12.203381
std	25.332252	15.224474
min	0.000000	0.000000
25%	1.000000	3.000000
50%	4.000000	8.000000
75%	13.000000	16.000000
max	846.000000	411.000000

Figure 1: The mean values of in-degree and out-degree are identical, the in-degree distribution has significantly higher variance. This seems that while authors cite a similar number of papers on average, a smaller subset of papers receive disproportionately more citations, creating a more skewed in distribution



3 Centrality Measures

Three centrality measures were used to assess the importance of papers in the citation network:

3.1 In-Degree Centrality

This metric simply counts the number of incoming links for each node, i.e., how many other papers cite a given paper. It is a naive yet intuitive measure of influence that 's often used in this field (e.g. Google Scholar).

3.2 Betweenness Centrality

This metric measures how often a node appears on the shortest paths between other nodes. It captures the notion of a node serving as a bridge in the network. Due to its computational cost, we used only a sample to estimate this measure for each node.

3.3 PageRank

This iterative algorithm assigns scores to nodes based on the probability of randomly landing on a node during a walk through the network. The specifics of PageRank-like algorithm have been discussed extensively in class. We used the teleportation variant with factor $\alpha = 0.85$, which as seen reflects the probability of continuing the random walk versus jumping to a random node, in order to prevent Spider Traps and Dead Ends. Regarding the TrustRank formulation, we avoided it as it comes with its own downsides (like having to choose trusted papers by hand and the centralization of the score) that overshadow its benefit, as papers are typically published on trusted media and checked by external experts.

Figure 2: Top 10 Papers by Each Centrality Metric

```
Top 10 papers by received citations:
Paper ID: 9803315, citations received: 846.000000
Paper ID: 9804398, citations received: 616.000000
Paper ID: 9407339, citations received: 557.000000
Paper ID: 9512380, citations received: 550.000000
Paper ID: 9606399, citations received: 542.000000
Paper ID: 9807344, citations received: 503.000000
Paper ID: 9306320, citations received: 464.000000
Paper ID: 9905221, citations received: 449.000000
Paper ID: 9408384, citations received: 444.000000
Paper ID: 9507378, citations received: 438.000000
```

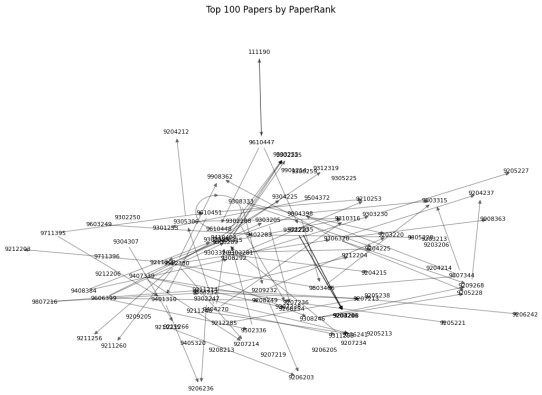
```
Top 10 papers by Paperrank:
Paper ID: 9303255, Paperrank Score: 0.001638
Paper ID: 9209205, Paperrank Score: 0.001265
Paper ID: 9310316, Paperrank Score: 0.001115
Paper ID: 9206203, Paperrank Score: 0.001035
Paper ID: 9208254, Paperrank Score: 0.000975
Paper ID: 9206242, Paperrank Score: 0.000854
Paper ID: 9803315, Paperrank Score: 0.000847
Paper ID: 9203203, Paperrank Score: 0.000835
Paper ID: 9303230, Paperrank Score: 0.000755
Paper ID: 9206236, Paperrank Score: 0.000726
```

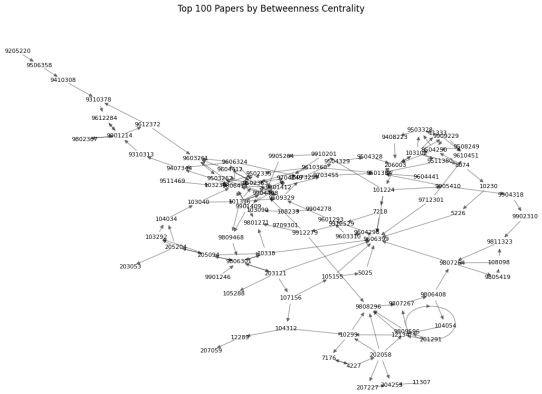
```
Top 10 papers by betweenness centrality:
Paper ID: 206003, Score: 0.153924
Paper ID: 101224, Score: 0.108573
Paper ID: 9501384, Score: 0.078588
Paper ID: 9806301, Score: 0.077856
Paper ID: 9806471, Score: 0.062899
Paper ID: 103230, Score: 0.061366
Paper ID: 9809468, Score: 0.056143
Paper ID: 9901409, Score: 0.053087
Paper ID: 9904408, Score: 0.049275
Paper ID: 9910201, Score: 0.048697
```

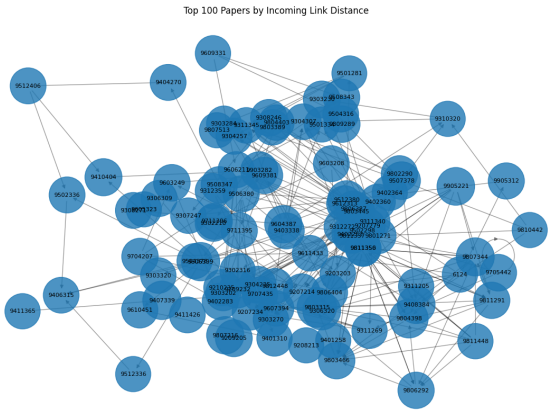
Figure 2: While both PageRank and incoming citations measure a paper's influence, they differ in scope: in-degree (citations received) counts all incoming links equally, while PageRank weighs citations from more influential papers higher and discounts those from sources that cite too broadly. This explains why some highly cited papers (e.g., ID 9803315) appear in both rankings, while others with high in-degree but less influential citers drop in the PageRank list. PageRank highlights prestige, not just popularity.

4 Top-100 Visualization

For a visual understanding of how influential papers are positioned in the graph, we plotted subgraphs induced by the top 100 nodes from each centrality measure.







5 Results

The comparison between the three ranking metrics reveals that while in-degree often captures raw popularity, PageRank refines this by considering the importance of citing papers. Betweenness centrality, though slower to compute, identifies papers that connect different parts of the network and might not be highly cited but are structurally significant.

We observed that some papers consistently appeared in the top 10 across all metrics, indicating both popularity and structural importance. Others appeared highly ranked in one measure but not others, highlighting differences in what each metric captures, potentially

6 Conclusion

We successfully implemented and compared three centrality metrics on a large real-world citation network. Each metric offers a different perspective on node importance, and their combined use provides a comprehensive understanding of the citation structure within the high energy physics literature.